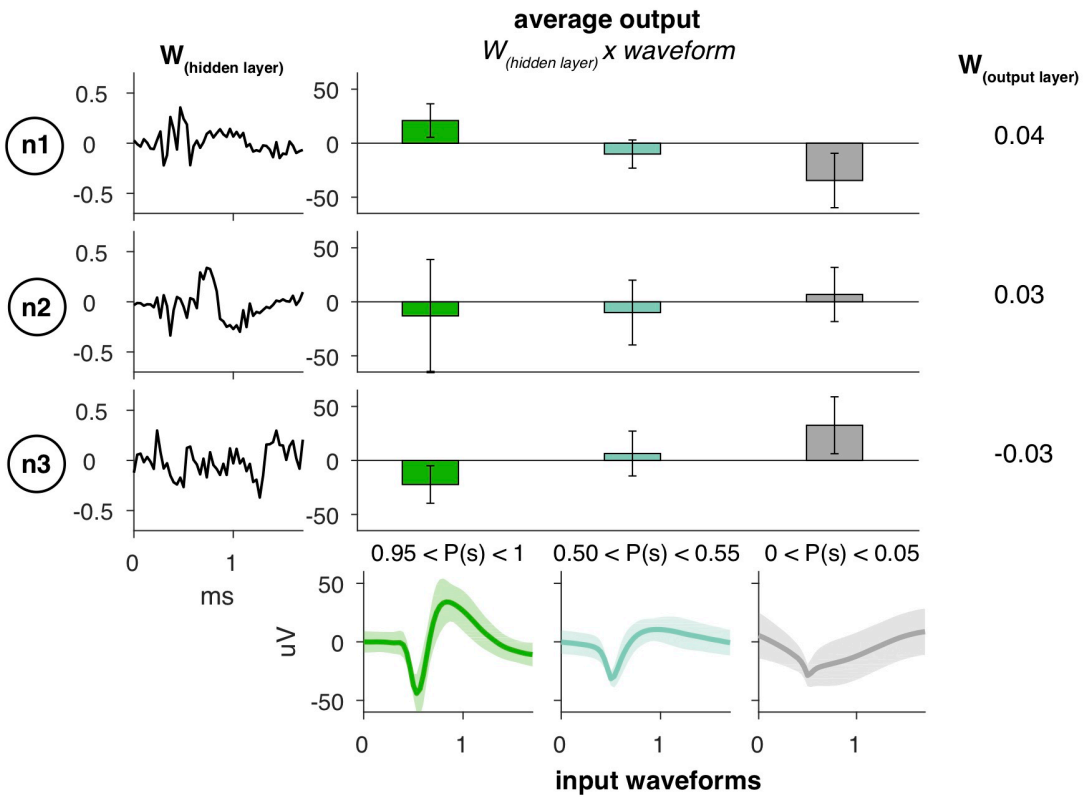
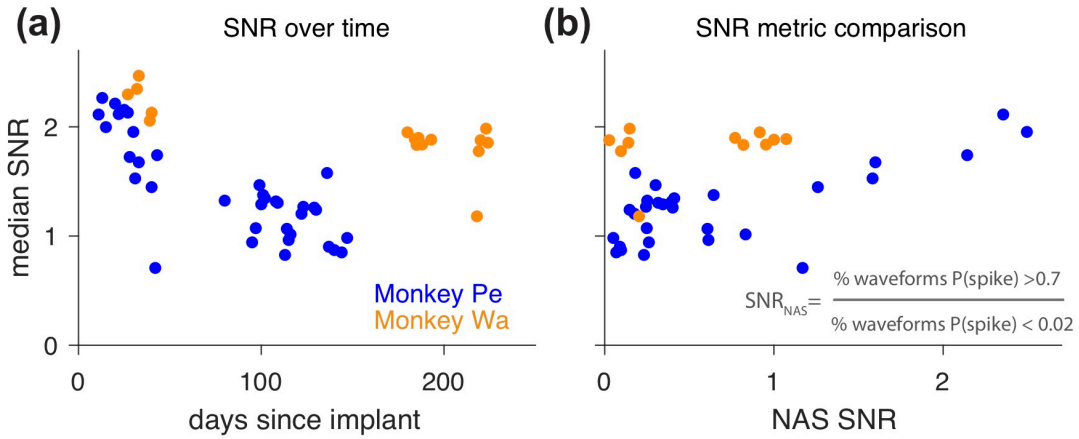
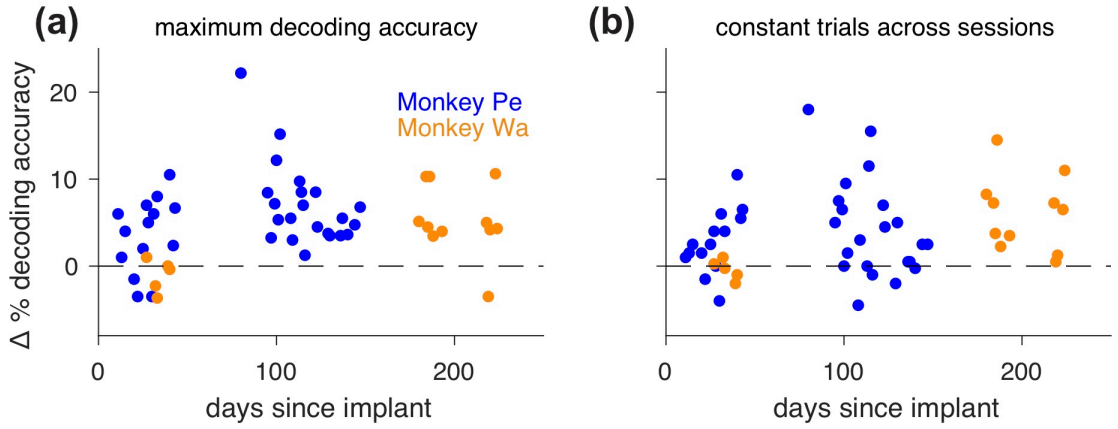


Supplementary Figure 1. Weights for three sample hidden units (n1, n2, n3) of the neural network hidden layer and their outputs. To demonstrate how the network classified waveforms, we used 1000 example input waveforms from each of three different categories (bottom three panels: mean and standard deviation of the waveforms in each category). The first category included very spike-like waveforms (i.e. those that were assigned a P(spike) value between 0.95 and 1 by the network, green), the second included moderately spike-like waveforms (0.50 < P(spike) < 0.55, blue), and the third included noise (0<P(spike)<0.05, grey). To visualize the hidden layer, we selected three out of the 50 hidden units and plotted their weights (left panels, $W_{\text{hidden layer}}$). Each waveform was linearly scaled by the weight vector of units n1, n2, and n3 and summed (middle panels: mean and standard deviation of the scalar outputs of $W_{\text{hidden layer}}$ x waveform; waveform category indicated by bar color and column, and unit # indicated by row). The output of the hidden layer was passed through a ReLU non-linearity before being multiplied by an output weight (right, $W_{\text{output layer}}$). The resultant values were summed and passed through a sigmoid non-linearity to produce a P(spike) value (not shown).

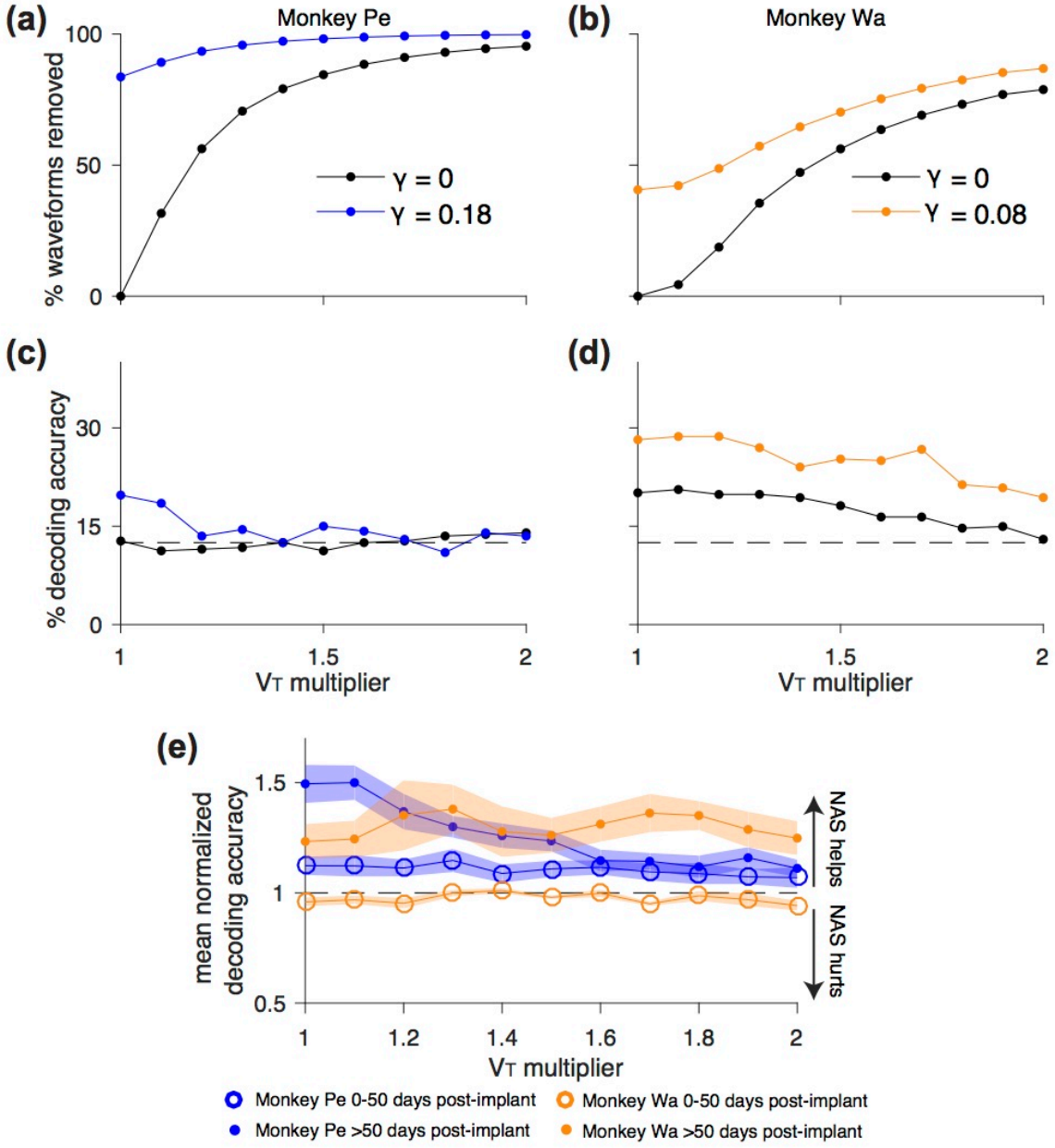
Overall, the network's hidden units were difficult to characterize; however, we observed that certain hidden units tended to have a larger, positive response to very spike-like waveforms (e.g. n1 green bar) while others had a larger, positive response to noise waveforms (e.g. n3 grey bar). The positive $W_{\text{output layer}}$ for n1 would then weigh the final summed output towards a P(spike) of 1 and the negative $W_{\text{output layer}}$ for n3 would weigh the output towards a P(spike) of 0. Unlike n1 and n3, n2 had variable responses to different types of waveforms, indicating that not every unit had a well-defined spike or noise preference. Overall, we found that different units in the hidden layer focused on features that were associated with different types of spikes in the training data. Note that in the network, the contribution of bias values was not shown in this visualization.



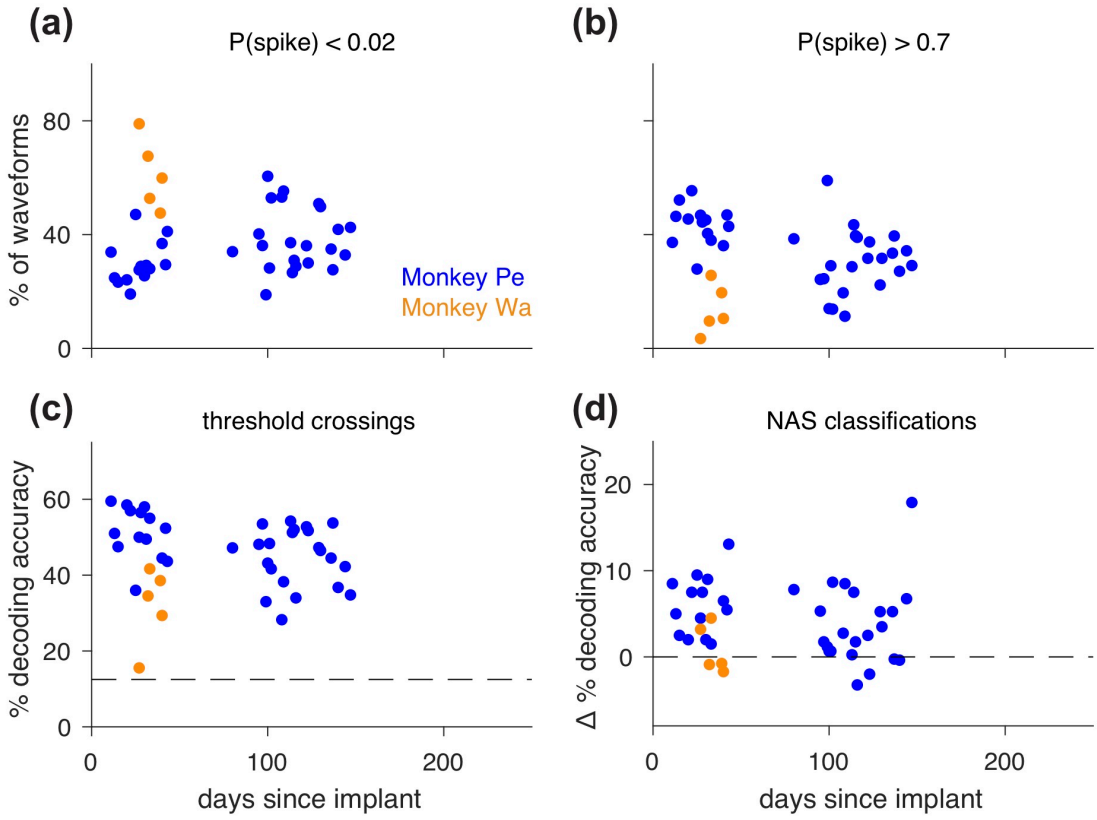
Supplementary Figure 2. SNR metrics decreased as the array aged for Monkey Pe (blue) and Monkey Wa (orange). **(a)** SNR was computed for each channel on the array using the method described in Kelly et al. (2007). The median SNR of the array decreased over time in both subjects (Spearman's rank-order, Monkey Pe: $\varrho = -0.77$, $p < 0.0001$; Monkey Wa: $\varrho = -0.72$, $p = 0.002$). **(b)** SNR_{NAS} (i.e. the ratio of the percentage of waveforms with a P(spike) > 0.70 to the percentage of waveforms with a P(spike) < 0.02) served as an alternate measure of SNR as it was correlated with median SNR over time (Spearman's rank-order, Monkey Pe: $\varrho = 0.78$, $p < 0.0001$; Monkey Wa: $\varrho = 0.75$, $p = 0.01$).



Supplementary Figure 3. Change in percent decoding accuracy from that with threshold crossings (Δ % decoding accuracy) over time, as in Figure 5 except **(a)** using the maximum decoding accuracy for each session and **(b)** keeping the number of trials used to decode constant across all sessions. In (a) the maximum decoding accuracy across all γ thresholds greater than 0 on each session was used to show the best possible network performance if γ was chosen separately in each session (and tested in a cross-validated fashion). Δ % decoding accuracy was significantly greater than zero for both subjects across all sessions (2-tailed, Wilcoxon signed rank test, Monkey Pe: $p < 0.0001$; Monkey Wa: $p = 0.01$). There was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Wa, but not in Monkey Pe (2-tailed, Wilcoxon rank sum, Monkey Pe: $p = 0.09$; Monkey Wa: $p = 0.01$). In (b), decoding accuracy was computed for each session with 25 trials for Monkey Pe and 50 trials for Monkey Wa. The γ threshold was selected using the same method described for Fig. 4d. Δ % decoding accuracy again was greater than zero for both subjects across all sessions (2-tailed, Wilcoxon signed rank test, Monkey Pe: $p < 0.0001$; Monkey Wa: $p = 0.002$). Again, there was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Wa, but not in Monkey Pe (2-tailed, Wilcoxon rank sum, Monkey Pe: $p = 0.68$; Monkey Wa $p = 0.001$).



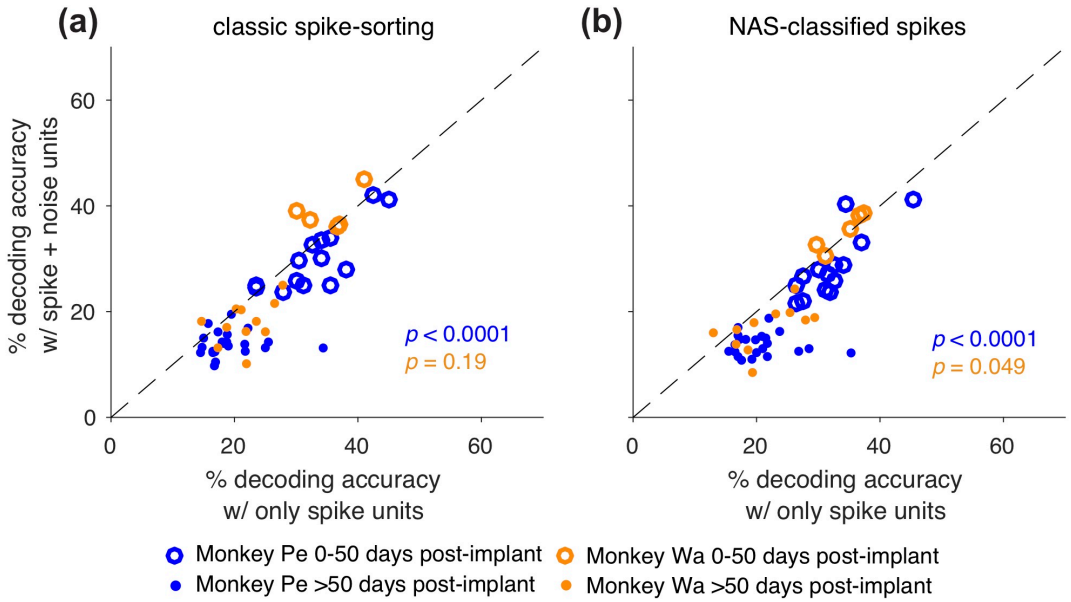
Supplementary Figure 4. Increasing the voltage threshold (V_T: the minimum voltage a waveform must cross to be included) did not improve decoding accuracy, but using the network classifications provided some benefit in most sessions. In lieu of increasing the RMS multiplier, which we did not store for our data, we increased the voltage thresholds (V_T) for each channel on a particular session by a factor between 1 and 2 (V_T multiplier). We only recorded threshold crossings in our data. A multiplier of 1 simply included all of our original threshold crossings, while a multiplier of 2 included only waveforms that crossed 2* V_T (the "new" threshold crossings). Waveforms that did not meet the new threshold were discarded. **(a,b)** Total percentage of waveforms removed at each V_T multiplier without using the network classifications ($\gamma = 0$, black line) and using the network classifications at a chosen γ threshold (blue line, Monkey Pe: $\gamma = 0.18$; orange line, Monkey Wa: $\gamma = 0.08$) for an example session in Monkey Pe and Monkey Wa. **(c,d)** Using the remaining waveforms after removing those that did not cross the new V_T, we computed decoding accuracy at each V_T multiplier without using the network classifications (black line) and after removing additional waveforms the network assigned P(spike) values below the chosen γ (blue and orange lines). Chance decoding was 12.5% (black dotted line). In both example sessions, increasing the V_T multiplier decreased decoding accuracy. **(e)** For each session, we normalized decoding accuracy with the chosen γ threshold (Monkey Pe: $\gamma = 0.18$; Monkey Wa: $\gamma = 0.08$) by the decoding accuracy without using the network classifications (i.e. the "new" threshold crossings) at each V_T multiplier. Then, we averaged across sessions recorded 0-50 days post-implant (early) and sessions recorded >50 days post-implant (late). Even at higher voltage thresholds, NAS improved decoding accuracy in both early and later sessions for Monkey Pe and in late sessions for Monkey Wa. In Monkey Wa's earlier sessions, using NAS did not substantially hurt or help decoding accuracy. These results were consistent with the trends observed in Figure 5.



Supplementary Figure 5. The network had a similar effect on decoding from V4 96-electrode Utah array implants in each animal as it did on decoding from the PFC array in the same sessions from the same animals (Monkey Pe- blue, Monkey Wa- orange). This analysis was done using the same steps as in Figure 5 except using waveforms from the V4 array instead of the PFC array. Monkey Wa did not have any usable late session (> 50 days post implant) data from the V4 array. As a result, only Monkey Pe was tested for trends over time.

(a) Percentage of waveforms with a P(spike) < 0.02 over time. A waveform with a P(spike) less than 0.02 was one that the network found very unlikely to be a spike. The percentage of these unlikely spike waveforms only slightly but significantly increased as Monkey Pe's array became older (Spearman's rank-order, $\varrho = 0.43$, $p = 0.01$). **(b)** Percentage of waveforms with a P(spike) > 0.70 over time. The percentage of waveforms that the network found to be strongly spike-like decreased as Monkey Pe's array became older ($\varrho = -0.48$, $p = 0.003$). The magnitude of these changes over time was much smaller in V4 compared to PFC. **(c)** Decoding accuracy with threshold crossings (i.e. $\gamma = 0$) decreased by a small amount as the array aged in Monkey Pe ($\varrho = -0.41$, $p = 0.01$). **(d)** Change in percent decoding accuracy using network classified spikes relative to decoding accuracy with threshold crossings (Δ % decoding accuracy). We computed a distribution of the maximum γ threshold (using the same method as in Figure 4c-d) for sessions that were 0-50 days post array implant and used the median to set the γ threshold before computing decoding accuracy for those sessions. We repeated this for sessions more than 50 days post array implant. In Monkey Pe, using NAS classifications improved decoding accuracy (2-tailed, Wilcoxon signed rank test, $p < 0.0001$) and in Monkey Wa, using the classifications neither hurt nor helped decoding significantly ($p = 0.81$). There was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Pe (2-tailed, Wilcoxon rank sum, $p = 0.0496$).

Thus, despite differences in the array quality between PFC and V4 (as demonstrated by the higher threshold crossing decoding accuracies in V4 compared to PFC in Fig. 5), using our network classifications still proved beneficial in one subject and had no significant impact on the other subject.



Supplementary Figure 6. Assigning waveforms classified as noise on each electrode to a new unit rather than discarding them did not improve decoding performance in most sessions for Monkey Pe (blue) or Monkey Wa (orange). In the main text analyses, any waveform classified as noise via our network or via spike-sorting was discarded. However, Todorova *et al.* (2014) found that creating a separate noise unit for the noise waveforms from each unit actually led to better decoding performance than discarding the noise. We sought to assess whether this result held for our data. **(a)** Using the manually spike-sorted classifications, we added a new unit to accompany each spike unit that contained all waveforms classified as noise (plotted on the y-axis). Percent decoding accuracy was worse with noise units for Monkey Pe (paired, 2-tailed, Wilcoxon signed rank test- $p < 0.0001$) and not significantly different for Monkey Wa ($p = 0.19$). **(b)** Similar to (a), except using NAS classifications. For each session, we discarded the noise units and found the γ threshold that resulted in the maximum decoding accuracy. We classified waveforms as spikes or noise using that threshold and then computed decoding accuracy with and without the noise units for each session (for the latter, the test set was independent from the set used to select the γ threshold). Discarding noise resulted in a higher decoding accuracy for most sessions in both subjects (Monkey Pe: $p < 0.0001$, Monkey W: $p = 0.049$). Thus, unlike Todorova *et al.* (2014), we did not find preserving the noise as separate units to be beneficial for decoding. In fact, the noise removed from most of the later recording sessions (>50 days post implant) was typically harmful for decoding.